Data Science Tools for Government Workers

Scott Hancock

2024-02-28

Contents

The latest version of this file can be found at its GitHub repository or downloaded here (right-click and "Save Link As…").

This is a collection of resources I've put together over the course of my career in government as an analyst. **Bold text** indicates that I've used it and found it very valuable; non-bold text means I perused it, thought it was interesting and worth saving for later.

The first two sections focus on the dominant data science languages, R and Python. These will contain many subsections on many similar topics, but will contain content specific to that language. The remaining sections are more generally applicable, or include resources for both languages.

Learning and using R

R is an open-source statistical programming language.

General overview / getting started

- Installing R on Windows, Mac OS X, and Ubuntu (DataCamp)
  - Note that "Step 1" and "Step 2" are mutually-exclusive
- Quick R
- **CRAN Task Views** (List of CRAN packages by subject/task area)
- Big Book of R (Oscar Baruffa's collection of R resources)

Digital textbooks and general reference

- **R for Data Science**
- R: Notes for Professionals
- Advanced R
- R for Journalists
- Univ of Cincinatti Business Analytics R Programming Guide
- An Introduction to R - Book written by ecologists
- Tidy Finance with R
- List of open source books about R

Training and Exercises

- Introduction to R (Free Datacamp course)
- R Exercises (Over 1000 exercises)
- swirl - an R package that teaches you how to use R, in the R console
- **Posit Webinars and Videos**
- Data Science in a Box (Data science course "in a box" using R, with learner and educator materials)

Example code and cheatsheets

- **Cookbook for R** Solutions to common tasks and problems in analyzing data.
- **RStudio's cheat sheet repository**

Importing data

- **How to connect R with Access database in 64-bit Windows?**
  - The solution from "Fiddler on the Roofies" seems best.
- **Using the tidyverse with Databases - Part I**

Data manipulation

- **Data Wrangling** (Altman, Behrman, Wickham)
- An Introduction to Data Cleaning with R (2013 discussion paper published by Statistics Netherlands)
- Flexible Imputation of Missing Data (Book by van Buuren, author of the {mice} package)
- **Demystifying Regular Expressions in R**

Learning and using Python

Configuring IDEs

- Setting Python Development Environment with VScode and Docker
- How to configure VS Code for AI, ML and MLOps development in Python

Books

- **Coding for Economists**
  - Python for Data Science is by the same author
- Python for Economists (Gallic) (321 page PDF)
- Introduction to Python for Econometrics, Statistics, and Numerical Analysis (Sheppard)
  - Book (PDF)
  - Course
- Computational and Inferential Thinking: The Foundations of Data Science (Book by Adhikari, DeNero, Wagner developed for UC Berkeley course Data 8: Foundations of Data Science)
- Python Programming for Data Science
- Tidy Finance with Python

Data visualization

Generally / multi-language

- Fundamentals of Data Visualization (Book) (Wilke)
- The Chartmaker Directory
- From Data to Viz - leads you to the most appropriate graph for your data. Code examples in Python, R, React, D3.js

Data visualization in R

- **The R Graph Gallery** (Website; usually the first place I go)
- Tufte in R
- Data Visualization in R with ggplot2 (Course website, Univ College London)
- A Curated List of Awesome ggplot2 Tutorials, Packages, Etc. (Website)

Data visualization in Python

- The Python Graph Gallery (Website; sister website to the R Graph Gallery)

Statistics and statistical techniques

- Library of Statistical Techniques (LOST)
  - "LOST is a publicly-editable website with the goal of making it easy to execute statistical techniques in statistical software."

- American Economic Association Continuing Education
  - 2017: Cross-Section Econometrics (AEA Webcast - Abadie, Angrist, Walters)
  - 2020: Mastering Mostly Harmless Econometrics (AEA Webcast - Abadie, Angrist, Walters)
- Introduction to Modern Statistics
- Linear Model and Extensions - (PDF book by Ding)
- Free statistics e-books for download (Blog post)

Research Design

- Research Design in the Social Sciences: Declaration, Diagnosis, and Redesign (Book by Blair, Coppock, Humphries)

Statistics in R

- Statistical Inference via Data Science: A ModernDive into R and the Tidyverse (Book by Chester Ismay and Albert Y. Kim)
- Modern Statistics with R (Book by Mans Thulin)
- Learning Statistics with R: A tutorial for psychology students and other beginners (Book by Danielle Navarro)
- The 9 concepts and formulas in probability that every data scientist should know (Blog post)
- ANOVA in R (Blog post)
- Advanced Data Analysis from an Elementary Point of View (Book by Shalizi)

Misc.

- The Elements of Statistical Learning (2009 book by Hastie, Tibshirani and Friedman)

Causal inference

Generally

- Causal Inference in Statistics: A Primer (PDF) (Pearl, et. al.)
- Brady Neal's Causality Blog (Computer Science PhD student; machine learning and "interventionist" focused)
  - Course (from 2020)
  - Book (PDF)
- An introduction to causal inference
- Causal Inference: What If? (2020 book by Hernán, Robins) (Includes code in R, Python, SAS, Stata, and other languages)
- **Causal Inference: The Mixtape** (Book by Cunningham, code in R, Python, Stata)
- **The Effect: An Introduction to Research Design and Causality** (Book by Huntington-Klein)
  - One of the gentler introductions
  - Nick Huntington-Klein's Youtube Page
- Impact Evaluation in Practice (World Bank handbook)
- Yale Applied Methods PhD Course
  - **Lectures**

        - ∗ I recommend only after familiarizing yourself with the topics

- A First Course in Causal Inference (PDF book by Ding)
- NBER 2021 Methods Lectures:
    - Causal Inference Using Synthetic Controls and Regression Discontinuity Designs
    - Rocio Titiunik, "Regression Discontinuity Designs: Foundations"
- Matteo Courthoud's "Awesome Causal Inference"
- Literature on Recent Advances in Applied Micro Methods (PDF; summaries and curation by Christine Cai)

Difference-in-Differences

- Taylor Wright's Youtube Playlists
    - DiD Reading Group
    - Other DiD Seminars

Causal inference in R

When a causal inference resource uses R exclusively, it appears here.

- **Causal Inference in R** (D'Agonstino-McGowan and Barrett)
    - A work in progress, but the "Asking Causal Questions" section is excellently done
- Research Design in the Social Sciences (book)
- Applied Causal Analysis (with R) (Book by Paul Bauer, intended to accompany a Univ of Mannheim course he teaches)
- Causal Analysis: Impact Evaluation and Causal Machine Learning with Applications in R (Martin Huber)

Causal inference in Python

When a causal inference resource uses Python exclusively, it appears here.

- **Causal Inference for the Brave and True** (Matheus Facure book)
    - Chapter 7, "Beyond Confounders", has probably the best discussion I've seen on which variables should and should not be part of a regression model.

Causal inference and machine learning

- Applied Causal Inference Powered by ML and AI (Book by V. Chernozhukov, C. Hansen, N. Kallus, M. Spindler, V. Syrgkanis)

Forecasting

- Forecasting: Principles and Practice - Book by Hyndman and Athanasopoulos; includes embedded videos
    - Hyndman's Monash University Applied Forecasting course

Machine Learning

Courses, videos, etc.

- An Introduction to Statistical Learning
  - Lecture videos
  - StanfordOnline: Statistical Learning with R (EdX)
    - "This is an introductory-level course in supervised learning, with a focus on regression and classification methods."
- Machine Learning and Econometrics - (Athey and Imbens, 2018 AEA Continuing Education Webcasts)
- Machine Learning & Causal Inference: A Short Course
  - Youtube playlist of 19 half-hour lectures by Susan Athey and others
- Applied Machine Learning with Tabular Data (Book by Kuhn & Johnson)
- Erasmus School of Economics: FinEML online seminar series

Machine Learning in R

- An Introduction to Statistical Learning with Applications in R (Free book)
- Hands-On Machine Learning with R

- TensorFlow for R
- Getting Started with Deep Learning in R

Machine Learning in Python

- An Introduction to Statistical Learning with Applications in Python

Data Mining

- Mining of Massive Datasets (Book by Leskovec, Rajaraman, Ullman)
  - The book is based on Stanford Computer Science course CS246: Mining Massive Datasets (and CS345A: Data Mining).

Geospatial

- US Census Cartographic Boundary Shapefiles
  - Consider the tidycensus package before using this

Washington State

- Univ of Washington: Geospatial Data Resources Guide: Washington State Geodata Resources
- **Washington State Parcel Database**
  - An incredible resource, but the page appears to be no longer maintained
- **Washington Master Addressing Services (WAMAS)**
- Washington Geospatial Open Data

Geospatial in R

- **Geocomputation with R** (Book by Lovelace, Nowosad, Muenchow)
- Spatial Data Science With Applications in R (Book by Pebesma (author of {sf}) and Bivand)
- Introduction to Visualizing Spatial Data in R (Website by Lovelace)
- Introduction to Spatial Analysis in R (Tutorials for the {sf} and {raster} packages)
- Mapping in R - Duke Univ. resource page for a workshop
- How to Create State and County Maps Easily in R (Blog post, Urban Institute)

Using Census Data in R

- Analyzing US Census Data: Methods, Maps, and Models with R (Book by Kyle Walker, creator of the {tidycensus} package)
- A Guide to Working with US Census Data in R (R Consortium)
- University of Michigan Institute for Social Research webinars:
  - Accessing and Analyzing US Census Data in R with Dr. Kyle Walker (3 hours)
  - Spatial Analysis of US Census Data in R with Dr. Kyle Walker (3 hours)

Other

- Lincoln Land Institute: Fiscally Standardized Cities

Project Management and Reproducible Research

Project and Data Management

- Data Management in Large-Scale Education Research (Book by Lewis)
- Veridical Data Science (Book by Yu and Barter)

Project Management and Reproducible Research in R

- **RMarkdown: The Definitive Guide** (Book by Xie, Allaire, Grolemund)
- **Happy Git and Github for the useR** (Book by Bryan, for STAT 545 course at Univ of British Columbia)
- A crash course in reproducible research in R (Blog post, 2016)
- Reproducible Research Using RMarkdown and Git through Rstudio (Tutorial, 2015)
- Google's R Style Guide

Reproducible Research in Python

- Google's Python Style Guide

Miscellanea

Text Mining in R

- Text Mining with R: A Tidy Approach (Book by Robinson and Silge, authors of the {tidytext} package)

Custom themes

- Custom themes in ggplot2 (Blog post)
- The Urban Institute's R theme

Other

- BERT - Basic Excel R Toolkit - Run R commands in Excel
- rOpenSci
- The rOpenSci Blog
- PyMC: Learn PyMC & Bayesian modeling (List of books and lessons for PyMC, a Bayesian modeling library in Python)
- Feature Engineering A-Z (Book by Hvitfeldt, using R and Python)

Washington State

- Washington State Auditor's Office - Financial Intelligence Tool
- Washington State Fiscal Information
- **Legislative Information Center: Documents and Publications**
  - A Citizen's Guide to the Washington State Capital Budget
  - A Citizen's Guide to the Washington State Budget
  - A Citizen's Guide to Washington State K-12 Finance
  - A Legislative Guide to Washington's Tax Structure
- OFM: A Guide to the Washington State Budget Process

Interesting blog posts

These blog posts were interesting or motivational to me as I learned R and later Python.

- (R) Why do we use arrow as an assignment operator?
- Setting up RStudio Server, Shiny Server, and PostgreSQL (Blog post) - Linux heavy, but was helpful for me when setting up a PostgreSQL server for a large study
- R in the data journalism workflow at FiveThirtyEight (Links to a video)
- How to Scrape Data from a JavaScript Website with R
  - Scraping Javascript-rendered web content using R
- NYT-style urban heat island maps
- Spatially Weighted Averages in R with sf
- Analyzing 1.1 Billion NYC Taxi and Uber Trips, with a Vengeance